Inverted Residual Bottlenecks with Large Kernel Attention for Remote Scene Classification

Russo Ashraf, Adri Priadana, Cao Ge, and Kanghyun Jo
Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, Korea
Email: ashrafrusso@gmail.com; priadana3202@mail.ulsan.ac.kr; caoge9706@gmail.com; acejo@ulsan.ac.kr

***Abstract*—Remote sensing image classification plays a pivotal role in environmental monitoring and urban planning, yet it faces the challenge of accurately interpreting complex and high-resolution images with fast inference speed for real time applications. To address this, we introduce the Mobile Large Kernel Attention Network (MLKANet), which integrates MobileNetV2's inverted residual structures with the large kernel attention mechanism from the Visual Attention Network (VAN). Our proposed MLKANet achieves a compelling balance of computational efficiency and sophisticated feature extraction, while maintaining the speed from the MobileNetV2 baseline. This study evaluates MLKANet's performance against state-of-the-art models using the Aerial Image Dataset (AID), demonstrating superior accuracy and efficiency. The architecture's effectiveness is further evidenced through an ablation study highlighting the scalability of our approach and class-wise performance analysis that showcases MLKANet's proficiency across various scene types.**

***Index Terms*—Remote Sensing, Scene Classification, MobileNetV2, Large Kernel Attention, Fast and Accurate Classification**

I. Introduction

Remote sensing has revolutionized the way we collect information about Earth's surface, making significant contributions to environmental science, urban planning, and disaster management. Yet, the classification of complex and high-resolution remote sensing images remains a formidable challenge. The heterogeneity and dynamic range of these images demand sophisticated analytical models capable of capturing both fine-grained details and broader contextual patterns. Deep learning models, particularly Convolutional Neural Networks (CNNs), have emerged as a game-changer in this field, automating the feature extraction process and providing robust representational capabilities. These models have significantly outperformed traditional machine learning approaches by learning hierarchical feature representations directly from the raw pixels of the images [1].

As remote sensing datasets increase in size and complexity, there is a growing need for models that can efficiently process vast amounts of data without compromising on performance. The MobileNetV2 [2] architecture presents a paradigm shift in this regard, offering a lightweight yet deep network design tailored for mobile and embedded vision applications. It achieves this by incorporating inverted residuals and linear bottlenecks, which efficiently manage the flow of information across the network while significantly reducing computational load. Its success in various vision tasks makes it an attractive backbone for resource-constrained remote sensing applications.

However, the nuanced spatial relationships and contextual dependencies characteristic of remote sensing images necessitate models that go beyond local receptive fields, prompting the integration of mechanisms that can capture long-range dependencies. Vision Transformers (ViTs) [3] have emerged as a powerful alternative to traditional CNNs, eschewing convolutional layers in favor of multihead attention mechanisms. This shift allows the model to learn long-range interactions between pixels, a critical factor for the intricate patterns often present in remote sensing images, but the quadratic complexity makes them impractical for high-resolution images. The Visual Attention Network (VAN) [4] takes a step in this direction by introducing the Large Kernel Attention (LKA) mechanism, which enhances the network's ability to understand and leverage the global context within an image. LKA offers a fine-grained, attention-driven approach to feature representation, crucial for handling the diversity and intricacies inherent in aerial scenes [5].

In this context, we introduce the Mobile Large Kernel Attention Network (MLKANet), which seamlessly blends the computational efficiency and depth of MobileNetV2 with the innovative LKA mechanism from the VAN framework. This hybrid model is engineered to optimize the classification of remote sensing images by capitalizing on the strengths of both architectures: the efficient information processing of MobileNetV2 and the sophisticated, global contextual understanding afforded by LKA. This approach is particularly apt for tackling the Aerial Image Dataset (AID), which encompasses a diverse array of scene types from various geographies, captured at different times and under varying conditions [6].

Our paper advances the state-of-the-art in remote sensing image classification with the following contributions:

- We introduce the MLKANet, a novel architecture that combines depth-wise separable convolutions with large kernel attention to enhance the feature extraction process for remote sensing images.
- Our comprehensive evaluation on the AID dataset not only demonstrates MLKANet's superior performance compared to established models but also showcases its effectiveness in dealing with high variability and complex spatial structures typical of remote sensing images.
- We compare with other state-of-the-art classification networks with similar speed and computational efficiency

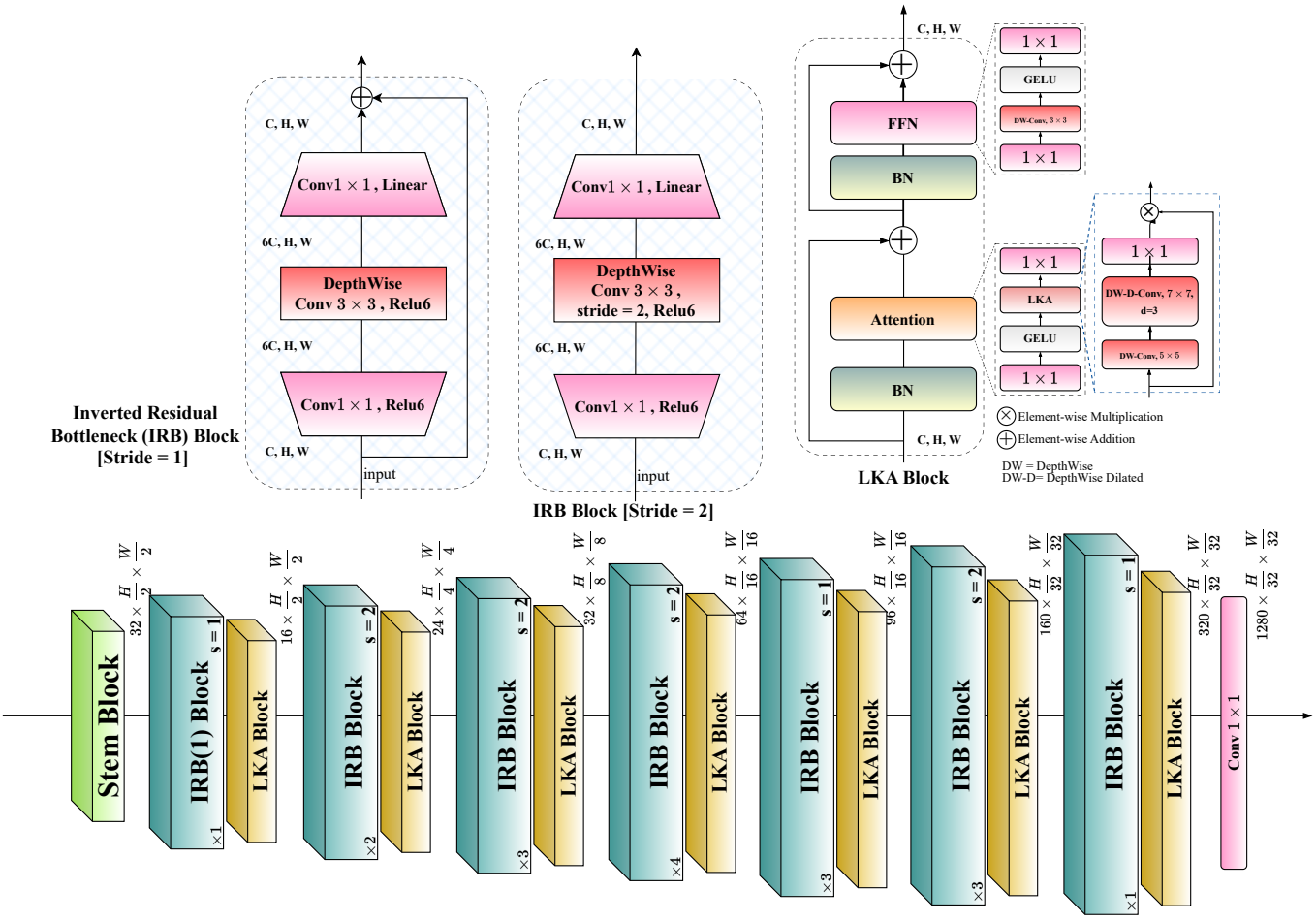

Fig. 1. Overall Architecture of the proposed Mobile Large Kernel Attention Network.

and showcase the strength of our architecture in both accuracy and efficiency.

II. Related Works

A. Recent Approaches for Remote Scene Classification

Recent developments in remote sensing scene classification have leveraged advanced deep learning techniques to address various challenges in the field, notably the management of high-dimensional data and the scarcity of labeled samples.One notable approach is the "Bag of Convolutional Features" (BoCF) method, which eschews traditional handcrafted features in favor of deep convolutional features, demonstrating superior effectiveness in scene classification [7]. Similarly, the "Bidirectional Adaptive Feature Fusion" strategy integrates features adaptively to enhance classification performance [8] Innovations also include task-specific models such as the "Task-specific contrastive learning for few-shot remote sensing image scene classification," which tailors the learning process to the unique characteristics of remote sensing data [9]. Authors in [10] proposes an efficient version of the popular densenet [11] architecture by significantly reducing the computation of the base model.

B. Towards Fast and Efficient Deep-Learning Models

The current quest in deep learning models revolves around achieving state-of-the-art accuracy while ensuring fast inference and computational efficiency. Convolutional Neural Networks (CNNs) have been the cornerstone of these advancements, where approaches like EfficientNet-B0 have systematically scaled up CNNs for improved performance across multiple tasks [12]. MobileNetV2 further streamlined the approach for mobile applications, optimizing the trade-off between accuracy and computational efficiency using inverted residuals and linear bottlenecks [2]. Attention mechanisms in CNNs, epitomized by the Visual Attention Network (VAN), have provided a means to model the global context within an image, a beneficial attribute for the complex spatial patterns encountered in aerial images [4]. Similarly, MobileViT-S introduced a fusion of convolutional principles with transformer-based architectures, aiming to marry the locality of CNNs with the global receptive field of transformers in a mobile-friendly package [13]. A novel addition to this suite of efficient architectures is FasterNet, which targets the reduction of floating-point operations (FLOPs) without a corresponding decrease in latency. By proposing a partial convolution

(PConv), FasterNet circumvents the inefficiencies of depthwise convolutions, significantly reducing both computational and memory overheads [14].

III. Methodology

A. Inverted Residual Bottleneck Block

The Inverted Residual Bottleneck Block, shown in Fig. 1 is a hallmark of the MobileNetV2 architecture, represents an approach to constructing deep neural networks that are both lightweight and computationally efficient. The block utilizes a two-step process: expansion and projection. Initially, the input is expanded to a high-dimensional space using a 1x1 convolution, allowing for a more expressive feature set. Subsequently, a depthwise separable convolution applies a 3x3 convolutional filter to each input channel separately, promoting efficiency. The final step involves projecting the features back to a low-dimensional representation using another 1x1 convolution, hence creating a 'bottleneck' that compacts the information while retaining the salient features. This design is pivotal for diminishing the model size and computational demand without significant losses in performance. The process within an IRB block can be mathematically described as follows:

$$f(x) = x + \text{Conv}_{1\times1}^{\text{Linear}}(\text{ReLU6}(\text{DWConv}_{3\times3} (\text{ReLU6}(\text{Conv}_{1\times1}^{\text{Expand}}(x))))), \quad (1)$$

where x represents the input to the block. The term $\text{Conv}_{1\times1}^{\text{Expand}}$ is the expansion convolution, $\text{DWConv}_{3\times3}$ denotes the depthwise convolution, and $\text{Conv}_{1\times1}^{\text{Linear}}$ is the projection convolution with a linear activation.

B. Large Kernel Attention Block

The Large Kernel Attention (LKA) block innovatively addresses the computational inefficiencies of traditional self-attention and large-kernel convolution methods used in computer vision. The LKA block, as depicted in Fig. 1, consists of a sequence of operations starting with an attention mechanism that assesses the input features to generate an attention map. This is followed by batch normalization (BN) to stabilize the learned features. The output from the BN is then passed through a feed-forward network (FFN), and the result undergoes another round of BN before being added to the original input through a residual connection, producing the final output of the block. By decomposing a large kernel into three components—a depthwise convolution for local spatial details, a depthwise dilated convolution for extended spatial reach, and a 11 convolution for channel adaptability—LKA captures both local and long-range dependencies effectively. The Attention operation in a LKA block is given by:

$$\text{Attention} = \text{Conv}_{1\times1}(\text{DW-D-Conv}(\text{DW-Conv}(F))), \quad (2)$$

$$\text{Output} = \text{Attention} \otimes F, \quad (3)$$

where F denotes the input feature map. This method enhances feature selection and noise reduction without the computational burden associated with standard convolutional approaches, facilitating dynamic adaptability in both spatial and channel dimensions, crucial for processing high-resolution images.

TABLE I
Performance Comparison of Classification Models

Model Name	Params. (M)	Flops (G)	Accuracy (%)	Inference Speed (FPS)	Model Size (MB)
MobileNetV2 [2]	6.83	2.26	90.85	293	18.4
EfficientNet-B0 [12]	6.87	4.06	90.45	291	32.7
Van-B0 [4]	6.92	3.85	88.7	289	31.1
MobileVit-S [13]	7.05	5.01	87.65	284	40.4
FasterNet-T2 [14]	6.95	13.74	90	288	110.1
MLKANet-0 (Ours)	6.95	4.01	**91.4**	288	32.8

TABLE II
Performance Comparison of Classification Models

Model Name	Params. (M)	Flops (G)	Accuracy (%)	Inference Speed (FPS)	Model Size (MB)
MLKANet-0	6.95	4.01	91.4	288	32.8
MLKANet-1	7.09	4.96	91.55	282	40.3
MLKANet-2	7.11	6.84	91.8	281	55.4

C. Overall Architecture of MLKANet

As shown in Fig. 1 the MLKANet architecture is conceived as a concatenation of efficiently designed blocks that together facilitate the accurate classification of remote sensing images. The network initiates with a stem block that conducts the initial convolutional processing to transform the input data into a feature-rich format suitable for the subsequent layers. Following this, a series of IRB blocks and LKA blocks are alternated, capitalizing on the efficiency of the former and the expansive contextual awareness of the latter. The alternation pattern allows the network to maintain a balance between local feature extraction and global context incorporation, which is particularly beneficial for capturing the diverse spatial relations in remote sensing images. The architecture concludes with a final convolutional layer that integrates the extracted features and prepares them for the classification task. Each block in the architecture is meticulously engineered to preserve the integrity of the image's spatial structure while ensuring the network remains computationally tractable.

IV. Experimental Analysis

A. Dataset and Implementation Detail

The Aerial Image Dataset (AID) is a large-scale collection of over 10,000 annotated 600x600 resolution RGB aerial images, encompassing 30 different classes representing various scene types. The AID's images were expertly labeled for remote sensing classification, featuring diverse scene types from various countries, primarily including China, the USA, England, France, Italy, Japan, and Germany. These images, consistent at a resolution of 600x600 pixels, vary in spatial resolution from 8 to 0.5 meters, captured across different seasons and under varying conditions.

TABLE III
CLASS-WISE PERFORMANCE OF VARIOUS EFFICIENT MODELS ON AID DATASET

Class Name	MobileNetV2	MLKANet-0 (Ours)	EfficientNet-B0	Van-B0	FasterNet-T2	MobileViT-S
Airport	87.59	87.22	**92.54**	88.89	85.51	85.93
BareLand	86.82	**90.77**	88.89	88.37	90.08	84.85
BaseballField	**98.80**	95.35	96.39	92.86	95.24	91.57
Beach	97.53	**98.16**	96.89	**98.16**	96.93	96.93
Bridge	**95.87**	94.40	90.32	90.16	86.40	89.08
Center	82.83	**86.60**	82.61	84.00	84.54	75.00
Church	88.89	**90.00**	88.89	83.95	88.31	85.00
Commercial	86.45	90.67	**91.39**	86.49	86.25	87.42
DenseResidential	92.12	**92.22**	92.02	90.36	92.12	91.86
Desert	90.78	**95.59**	91.30	91.30	92.75	91.97
Farmland	**93.17**	**93.17**	92.22	88.34	90.57	83.95
Forest	**99.03**	98.08	98.08	96.15	98.08	96.15
Industrial	**94.79**	76.03	79.35	78.89	84.21	75.15
Meadow	94.61	**98.39**	96.77	95.16	98.36	98.39
MediumResidential	**97.53**	94.74	94.55	92.86	92.98	89.66
Mountain	94.51	**96.40**	87.50	94.81	95.77	90.51
Park	**96.36**	80.56	80.26	81.12	83.69	84.44
Parking	96.45	**99.38**	97.44	96.15	96.77	99.37
Playground	**97.53**	92.21	95.42	91.50	94.19	93.51
Pond	94.51	95.08	93.99	91.80	94.38	**95.56**
Port	96.51	96.47	93.49	94.67	95.35	94.92
RailwayStation	**92.73**	92.04	91.59	90.91	82.69	87.27
Resort	71.29	**76.29**	76.00	74.75	72.92	57.43
River	93.62	**95.59**	91.97	90.51	92.31	86.76
School	56.57	58.49	71.03	57.66	**72.88**	61.82
SparseResidential	98.25	**100.00**	99.12	96.43	95.58	93.91
Square	**82.54**	81.36	78.63	69.42	78.40	67.86
Stadium	94.66	94.03	94.66	**95.45**	93.13	92.19
StorageTanks	92.42	**93.33**	89.39	85.71	88.55	88.57
Viaduct	93.51	**98.06**	94.27	92.99	90.68	90.57

To ensure consistency across experiments, all models were trained from scratch using the same test split, with input images resized to 224x224 pixels and data augmentation techniques such as AutoAugment and random flips applied. We split the dataset into train set, validation set and test set using a 60:20:20 ratio. The LION optimizer was used for its efficiency, along with the standard Cross-Entropy Loss function. Learning rates were set lower for Transformer and Hybrid models, and all models were trained on NVIDIA Tesla V-100 GPU, with training epochs set to 100.

B. Evaluation on AID Dataset

The detailed results on AID dataset is shown in Table I along with the comparative analysis with the other state of the art methods. Our proposed, MLKANet-0 demonstrates enhanced classification performance with a 91.4% accuracy, outperforming established models like MobileNetV2, EfficientNet-B0, Van-B0, and MobileViT-S. It accomplishes this with a computational requirement of 4.01 gigaflops, positioning it as a more efficient alternative to the particularly compute-intensive FasterNet-T2, which demands 13.74 gigaflops. Despite this efficiency, MLKANet-0 retains a swift inference speed of 288 frames per second, on par with other models and notably more efficient than the slightly slower MobileVit-S. In terms of size, MLKANet-0 occupies 32.8 MB, comparable to Van-B0 and EfficientNet-B0, and is substantially smaller than the larger FasterNet-T2, which occupies 110.1 MB. This balance of accuracy, efficiency, and model size underscores MLKANet's suitability for applications where model compactness and performance are critical. The proposed model outperforms the other recent methods, while not sacrificing speed.

C. Ablation study of MLKANet Variations

Table II shows the study examining variations of the MLKANet architecture, performance enhancements are systematically evaluated in relation to increments in model complexity. The study specifically explores the impact of adjusting the MLP ratio within the LKA Block. The base model, MLKANet-0, with an MLP ratio r=2, demonstrates solid performance with an accuracy of 91.4% and the fastest inference speed among the variants at 288 FPS, all contained within a 32.8 MB model size. Advancing to MLKANet-1, where the MLP ratio is doubled to r=4, a slight increase in accuracy to 91.55% is observed, alongside a modest increase in model size to 40.3 MB and a negligible decrease in inference speed. The most substantial model, MLKANet-2, with an MLP ratio of r=8, achieves the highest accuracy of 91.8%, at the cost of further increased computational complexity, leading to a larger model size of 55.4 MB and a minimally slower inference speed of 281 FPS. This stratified augmentation underscores the trade-off between model size and performance within the MLKANet series. But, even at the highest size MLKANet-2 retains fast inference speed while improving the accuracy by almost 1% over the baseline MobileNetV2.

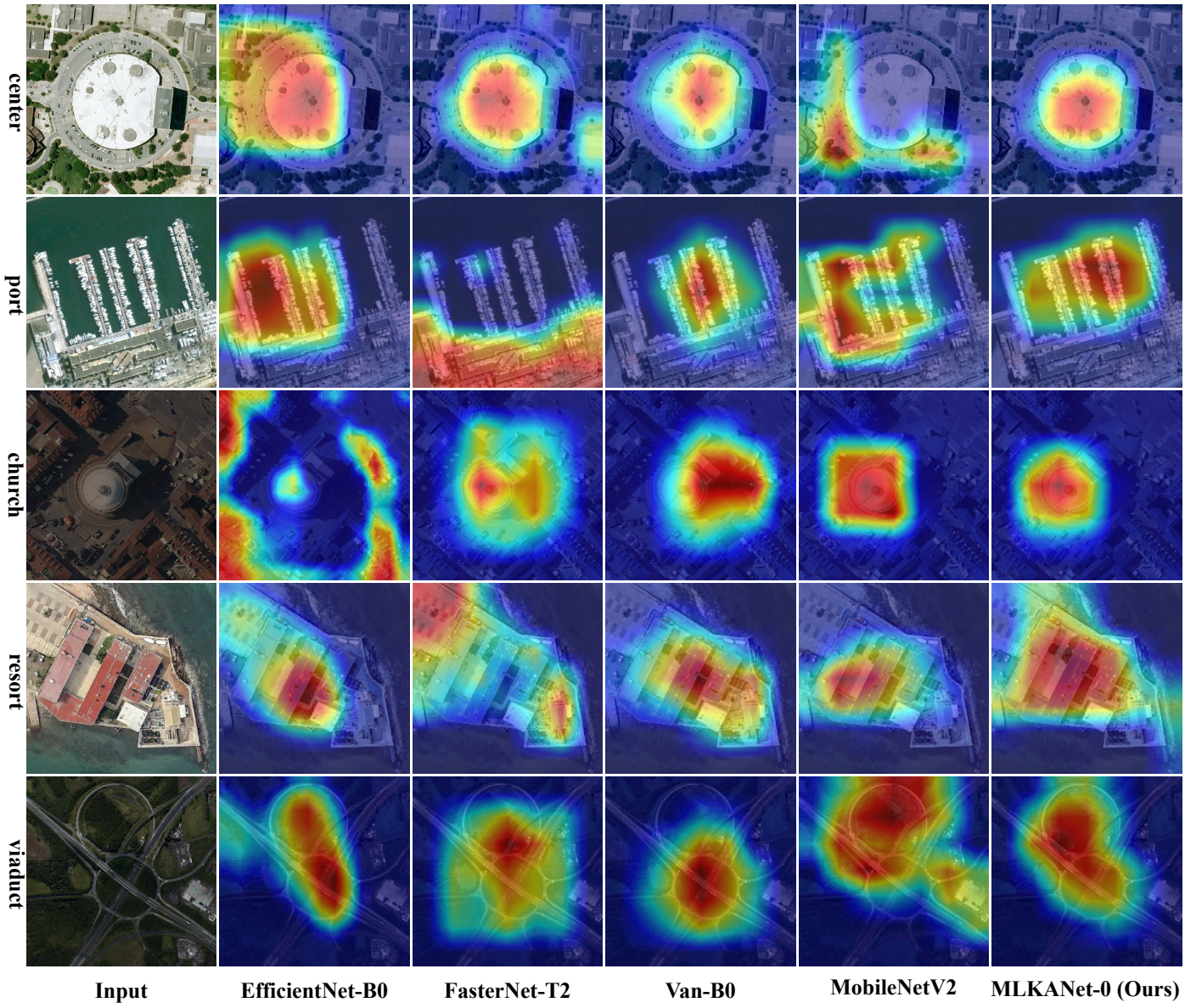

Fig. 2. Class activation map from 5 example classes from the AID dataset is shown using Eigen-Cam [15] algorithm, our model MLKANet-0 is compared with EfficientNet-B0, FasterNet-T2, Van-B0 and MobileNetV2.

D. Class-Wise Performance

Shown in Table. III the class-wise performance for the AID dataset reveal that MLKANet-0 achieves the highest accuracy in 15 out of 30 classes when compared to other efficient models like MobileNetV2, EfficientNet-B0, Van-B0, FasterNet-T2, and MobileViT-S. This indicates MLKANet-0's superior capability in accurately distinguishing a broad range of aerial scene types, from natural landscapes like 'Forest' and 'Farmland' to constructed areas such as 'Airport' and 'Parking', substantiating its effectiveness for diverse remote sensing image classification tasks.

E. Qualitative Analysis

Fig. 2 shows the Eigen-Cam [15] activation maps of 5 classes center, port, church, resort and viaduct from the AID dataset and compares the models of Table I. The provided Eigen-CAM activation maps for various models using AID dataset samples showcase the distinct focus areas that each model deems significant when classifying scenes. The MLKANet-0, our proposed network, consistently exhibits more targeted and central activation patterns, particularly on salient features of the scene types, indicating a better alignment with discriminative regions. This focused activation is indicative of the model's robust feature extraction capabilities, which likely contributes to its superior performance in scene classification tasks as compared with EfficientNet-B0,

FasterNet-T2, Van-B0, and MobileNetV2.

V. Conclusion

This paper presented MLKANet, an architecture adept at handling the intricate task of remote sensing scene classification. By effectively combining depth-wise separable convolutions and large kernel attention, MLKANet not only excels in recognizing diverse scene types but does so with impressive computational economy. Through extensive experiments on the AID dataset, MLKANet outperformed comparable models, achieving the best results while maintaining comparable inference speed with MobileNetV2. The ablation study confirmed that increasing the MLP ratio within the LKA Block correlates with incremental accuracy gains. Additionally, our qualitative analysis, demonstrated by activation maps, reaffirmed the network's focused and discerning feature extraction capabilities. MLKANet stands out as a significant contribution to the field of remote sensing, offering an efficient and robust tool for high-fidelity image classification.

Acknowledgment

This result is supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003)

References

[1] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, pp. 733–752, 2023.

[5] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[6] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[7] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1735–1739, 2017.

[8] X. Lu, W. Ji, X. Li, and X. Zheng, "Bidirectional adaptive feature fusion for remote sensing scene classification," *Neurocomputing*, vol. 328, pp. 135–146, 2019.

[9] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 143–154, 2022.

[10] R. M. A. Uddin, T.-D. Tran, G. Cao, and K.-H. Jo, "Densenetx: Efficient densenets for remote scene classification without pretraining," in *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*. IEEE, 2023, pp. 1–6.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[13] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[14] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 021–12 031.

[15] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–7.